National College of Ireland

Configuration Manual

MSc Industrial Internship
Cyber Security

Kevin Shaji
Student ID: x22108718

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Kevin Shaji……

Student ID: …… x22108718……

Programme: ……MSc Cyber Security…… **Year:** …2023…..

Module: ……Industrial Internship……

Supervisor: ……Vikas Sahni……

Submission Due Date: ……January 5th 2024……

Project Title: ……Privacy Impact Assessment of Third-Party Dependencies……..

Word Count: ……1355…… **Page Count**……12……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Kevin Shaji……

Date: ……January 5th 2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Kevin Shaji
Student ID: x22108718

1 Introduction

This configuration handbook contains the fundamental setup and equipment's used to complete this project work for Cubic Telecom[1]. This project intends to create a framework that uses Natural Language Processing (NLP) techniques learning algorithms to conduct a Privacy Impact Assessment over third party dependencies privacy policies and provide insights over each analysed dependencies and generate impact scores for each of them. The configuration handbook is crucial and will include all the necessary software, hardware, and implementation techniques to develop this project.

2 Hardware Requirements

Operating System: Windows 11
RAM: 16.0 GB
Processor: 11th Gen Intel Core i5-11320H @ 3.20GHz 2.50 GHz
Storage: 512 GB SSD
System Type: 64-bit operating system, x64-based processor

3 Software Requirements

Python 3.9.6: Python was the core language used for the development of this framework.

Flask 3.0.0: Flask framework[2] was utilized to develop an API service over the project to handle user requests and host the application locally.

Requests 2.7.0: It is a python module used to send HTTP requests and fetch responses in the program.

Beautiful Soup 4.12.2[3]: It is used to scrap various privacy policy documents for creating the dataset.

Google search 1.2.3: It is a python module used to find out the privacy policy documents and fetch the links from online resources utilizing the google search api.

Jsonify 0.5[4]: A python module utilized to convert the data into json format while processing the requests.

[1] Cubic Telecom: https://www.cubictelecom.com
[2] Flask: https://flask.palletsprojects.com/en/3.0.x/
[3] Bs4: https://pypi.org/project/beautifulsoup4/
[4] Jsonify: https://pypi.org/project/jsonify/

Postman v10: Postman[5] is a tool used to test the application programming interface by crafting appropriate request methods and data. The tool was used to test the developed API endpoint of the framework.

SQLite3: It was utilized to store the details of the analysed data including the privacy insights and the scores in the local machine for proper testing of the developed framework.

Joblib 1.3.2: Joblib[6] is a Python library primarily used for lightweight pipelining in Python. It is particularly useful for efficiently parallelizing and caching functions that involve computation-heavy tasks. The NLP model was converted into a Joblib file for smooth functioning of the framework.

4 Building the Privacy Impact Assessment framework

Building a novel PIA framework requires the use of Natural Language Processing (NLP) capabilities including sentimental analysis to build a machine learning model that produces valuable insights with regarding to the privacy of a third-party dependency after analysing the policy documents. BERT (Bidirectional Encoder Representations from Transformers)[7] is a powerful pre-trained NLP model by Google. BERT(BERT Explained: State of the art language model for NLP | by Rani Horev | Towards Data Science) belongs to the transformer architecture and is designed to understand context and relationships between words in a sentence. This capability of BERT is leveraged in the training of the PIA model for the framework. It has achieved state-of-the-art performance in various NLP tasks, such as text classification, named entity recognition.

The dataset was scrapped from various privacy policy documents of third-party dependencies using Python and libraries such as Beautiful Soup, the data is annotated using the metrics provided by the Cubic Telecom to meet their needs.

PIA framework was developed using core python and libraries and it can run as a command line tool, Flask framework was used to run it as an API service.

5 Installing Dependencies

Installing the python dependencies using the following command

```
pip3 install requests
pip3 install bs4
pip3 install joblib
pip3 install google
pip3 install jsonify
pip3 install joblib
```

[5] Postman: https://www.postman.com/downloads/
[6] Joblib: https://pypi.org/project/joblib/
[7] https://www.turing.com/kb/how-bert-nlp-optimization-model-works

It is assumed that the artifacts have been downloaded and extracted into a folder. File Structure overview is listed below –

- main.py – Contains the core code of the application imported all necessary modules including the model the database connections for running the application as command line tool directly from the terminal.

- model.py – Contains the NLP model exported as a Joblib file: 'pia_analyser.joblib' and the required functions to perform the privacy policy analysis, appropriate comments are present to explain each block of code.

- database.py – Contains the sqlite3 database connection programs and the insert and retrieval commands for each analysed third party dependencies. It is not necessary to install sqlite3 module. It is included in the standard library (since Python 2.5). After the initial execution of the program a database named 'pia-database.db' is created in the same project directory.

- Scrapper.py – Contains all the necessary code to search and fetch the privacy policy document from the web. It is using google search module in python to conduct an option is displayed to enter the privacy policy url manually.

- Server.py – Contains all the code written in flask framework to provide the service as an api endpoint. By default, it is served using flasks local server at 127.0.0.1:5000. Ensure that port is free without running any other services. Tools like postman can be used to fetch and provide requests to the program.

6 Running the program

Before executing the program, ensure that you have downloaded all artifacts, implemented the necessary changes as outlined in the Section 5, and installed all required dependencies.

Running program as a command line tool:

Step 1: Open the terminal in the project folder.

Step 2: Run the main program 'python3 main.py'

Figure 1: PIA tool command line interface.

Step 3: Choose the required option, here we will start with analysing a new dependency and provide the dependency name select option 2:

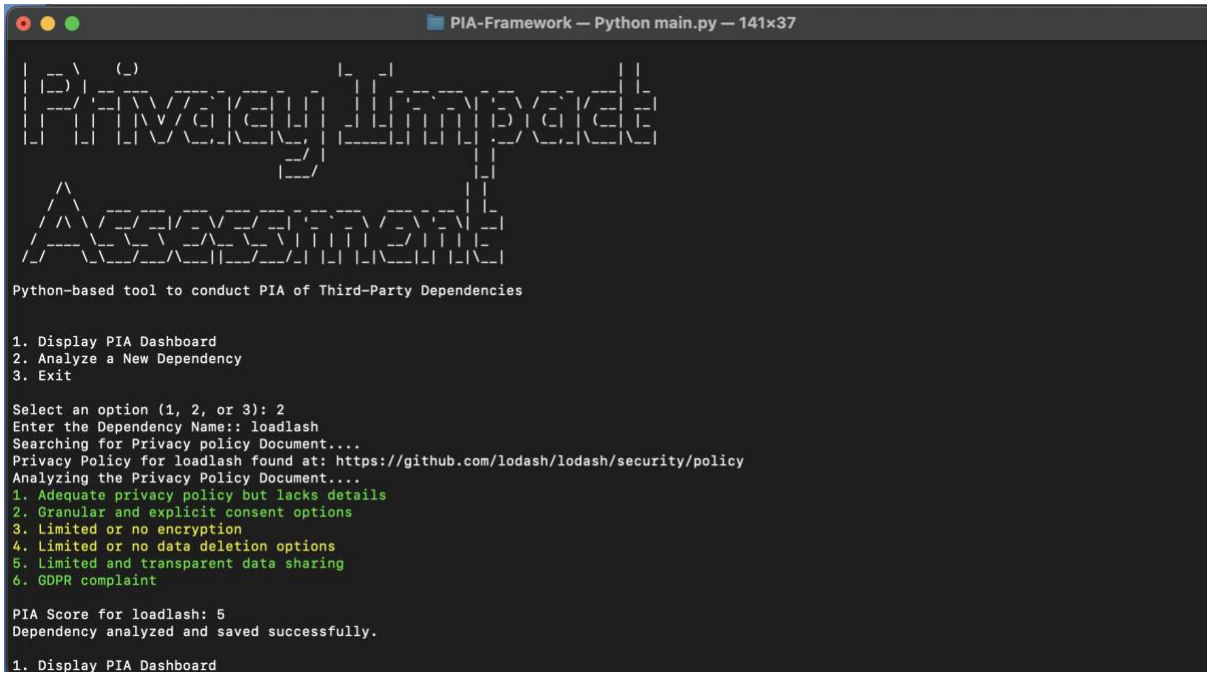

Figure 2: Analysing a privacy document and providing Insigts and PIA score.

Step 4: The program has run successfully on the selected dependency and provide the results and saved it to the database. To fetch the dashboard of the analysed dependencies, select option 1:

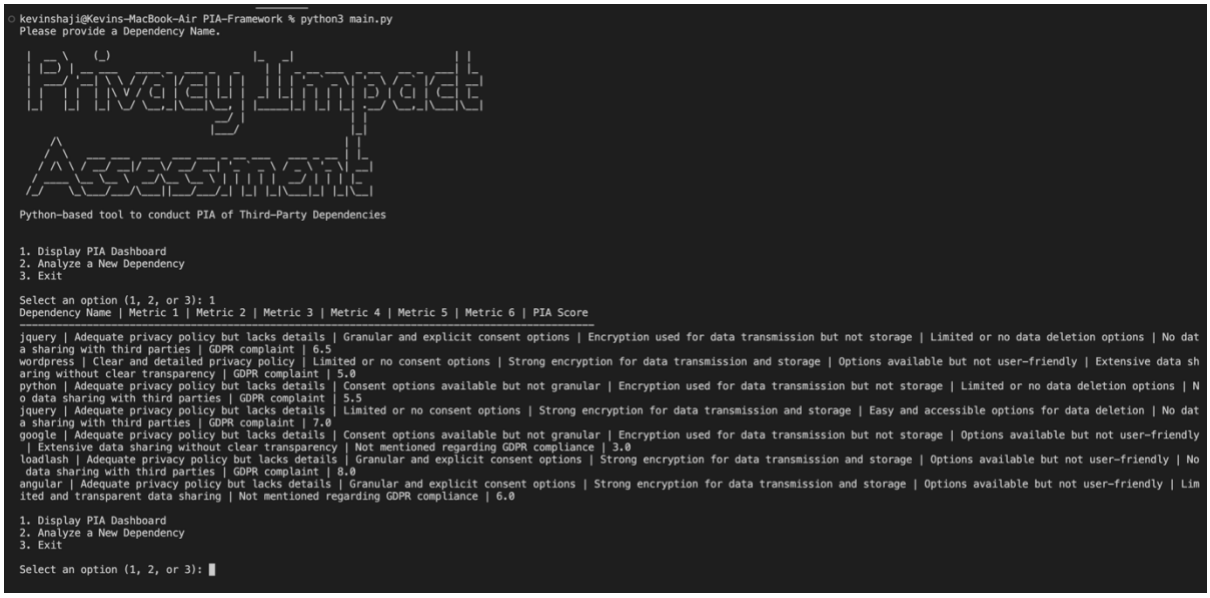

Figure 3: PIA tool command line dashboard display of results.

Running the program as an API service endpoint:

Step 1: Browse to the program directory and run the command ‘python3 server.py’.

Figure 4: PIA tool serving as an API in the local system.

Step 2: Open Postman application and browse towards http://127.0.0.1:6000/pia_dasboard to generate a GET request to fetch all the analysed dependency results in a JSON format:

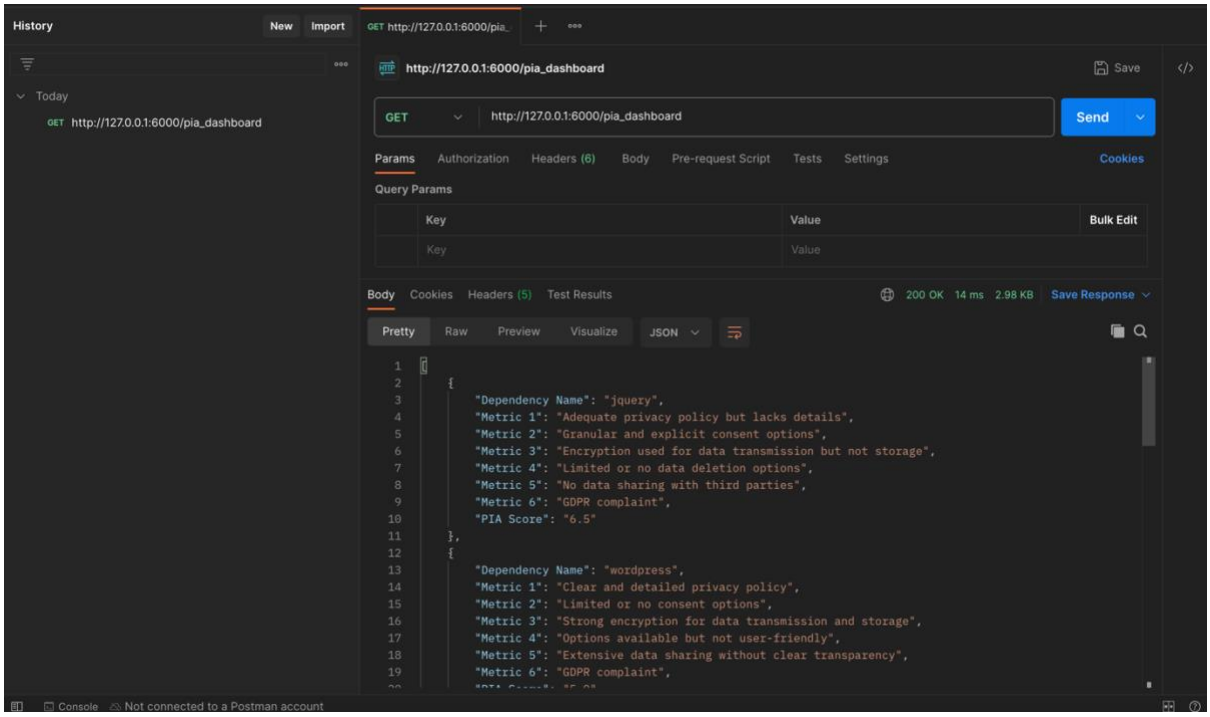

Figure 5: Using Postman to fetch the dashboard data.

Step 3: To analyse a new dependency via an API request. Send a POST request to http://127.0.0.1:6000/analyze_dependency/ set the header value as 'Content-Length: application/json' as the server accepts only json requests:

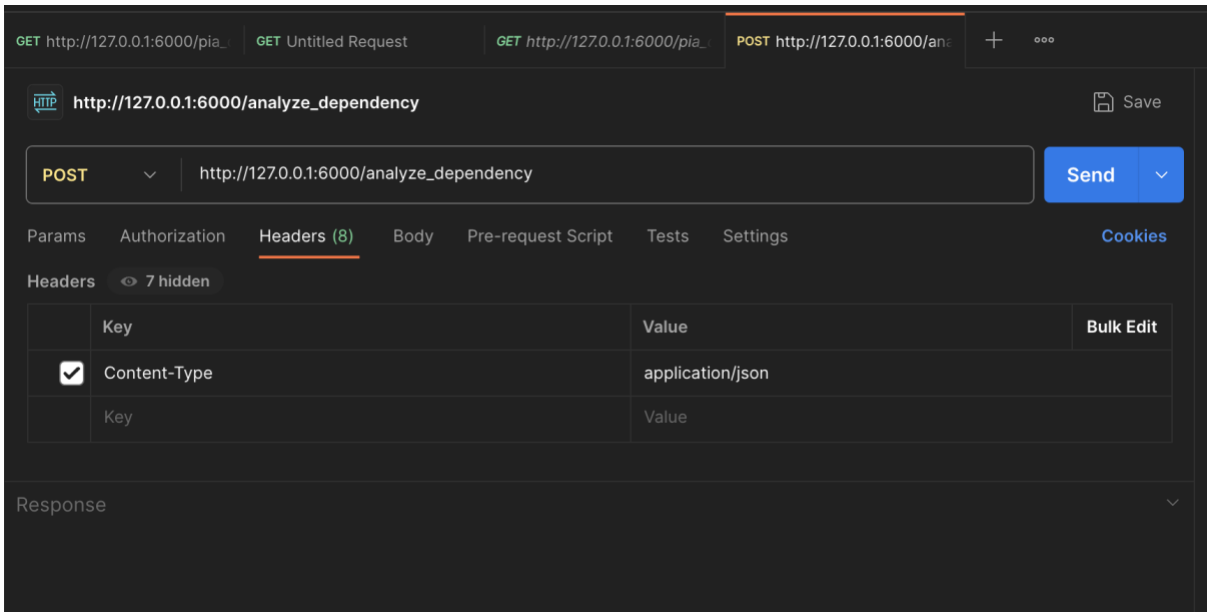

Figure 6: Configuring Postman to send the request body for analysis.

Step 4: Send the POST request in the following JSON format including the dependency name to conduct analysis:

```
{
   "dependency_name":"wordpress"
}
```

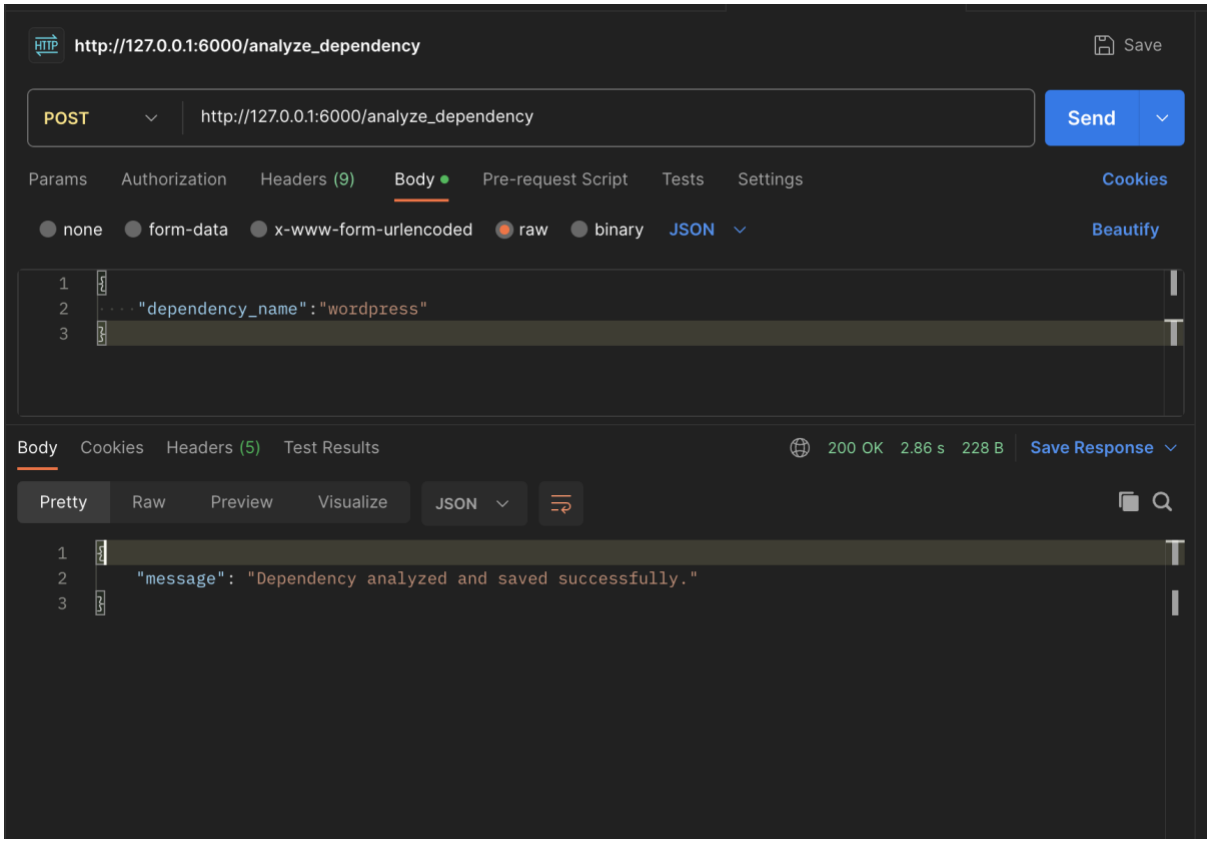

Figure 7: Analysing third party dependencies using data embedded as a JSON body.

References

BERT Explained: State of the art language model for NLP | by Rani Horev | Towards Data Science (no date). Available at: *https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270* (Accessed: 8 December 2023).

Appendix H – Monthly Internship Activity Report

The Internship Activity Report is a 1-page monthly summary of the activities performed by you and what you have learned during that month. The Internship Activity Report must be signed off by your Company and included in the configuration manual as part of the portfolio submission.

Student Name: Kevin Shaji Student number: x22108718

Company: Cubic Telecom Month Commencing: September

This month was spent on performing the Literature review on the Privacy Impact Assessment in third party dependencies.

Subsequently, the methodology to be utilized in this research was determined.
Various models were scrutinized before reaching a final decision.

Conducted penetration testing on multiple API and web applications within the organization's defined scope.

Employer comments

Student Signature: _____ Date: 20/12/2023

Industry Supervisor Signature: _____ Date: 20/12/2023

Appendix H – Monthly Internship Activity Report

The Internship Activity Report is a 1-page monthly summary of the activities performed by you and what you have learned during that month. The Internship Activity Report must be signed off by your Company and included in the configuration manual as part of the portfolio submission.

Student Name: Kevin Shaji Student number: x22108718

Company: Cubic Telecom Month Commencing: October

Manual annotation of the dataset and tokenisation was performed during this month and defined the criteria and the insights outputs including the metric scores for the privacy impact assessment framework.

Conducted comprehensive wireless penetration testing activities within the established infrastructure of the company, thorough assessments and analyses to ensure the security and resilience of the wireless network.

Employer comments

Student Signature: _____ Date: 20/12/2023

Industry Supervisor Signature: _____ Date: 20/12/2023

Appendix H – Monthly Internship Activity Report

The Internship Activity Report is a 1-page monthly summary of the activities performed by you and what you have learned during that month. The Internship Activity Report must be signed off by your Company and included in the configuration manual as part of the portfolio submission.

Student Name: Kevin Shaji Student number: x22108718

Company: Cubic Telecom Month Commencing: November

Training and validation the developed model with various privacy policy documents and redefining the metrics. Integrated the model into a command line tool written in python and deployed it as an API service which can be used to fetch the assessment results into vulnerability management dashboards

Conducted Mobile penetration testing activities for both Android and iOS devices including static and dynamic testing.

Employer comments

Student Signature: _____ Date: 20/12/2023

Industry Supervisor Signature: _____ Date: 20/12/2023